Issues in Spoken Dialogue Systems: Experiences with the Dutch ARISE System

Janienke Sturm[1], Els den Os[2], Lou Boves[1,2]

[1]University of Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
[2]KPN Research, P.O. Box 421, 2260 AK Leidschendam, The Netherlands
sturm@lands.let.kun.nl, e.a.denos@research.kpn.com, boves@lands.let.kun.nl

ABSTRACT

In the ARISE project we developed an experimental spoken dialogue system for access to train timetable information of the Dutch Railways. It is based on an existing system (VIOS), which became operational during the course of the project. This paper discusses a number of issues that came to light during evaluations of the ARISE system. We are able to shorten the dialogue with the system considerably, by using confidence measures, short prompts and a zooming exceptions handling strategy. The problems that we observed are all related to the fact that both the dialogue system and the user are unable to infer each other's intentions and capabilities. Also, evaluation of a dialogue system is difficult since the evaluator has little means to infer the real intentions of the user. Working with predefined scenarios is not the optimal solution for this problem.

1 INTRODUCTION

Compared to the number of laboratory studies, few experiments on spoken dialogue information systems have addressed interactions between paying customers and commercially deployed systems. Given the enormous logistic and methodological problems involved in investigations of operational systems, this state of affairs is not surprising. As a consequence, design and implementation decisions for operational systems are necessarily based on extrapolations from laboratory tests.

In the LE-3 project ARISE (Automatic Railway Information Systems in Europe) an experimental spoken dialogue system for access to timetable information of the Dutch Railways has been built. It is closely related to an existing system (VIOS) that became operational in the Netherlands during the course of the project. In ARISE we could compare the functionality and dialogue structure of the laboratory system with the operational system.

Two versions of the ARISE system were tested; we will call these the 98-system and the 99-system. In [6] a test was reported that compared the 98-ARISE system with the VIOS system. In [4] we reported on an evaluation of the same 98-system; subjective as well as objective performance measures were obtained with 68 naive subjects who had to perform three scenarios of increasing difficulty. Recently, we completed a similar experiment with the 99-system, using expert judges in addition to naive users [5]. For both tests the same scenarios were used.

In this paper we take the ARISE and VIOS systems as the point of departure to discuss issues in spoken dialogue systems of which we think that they also apply to other information applications.

2 IMPLICIT *VS.* EXPLICIT CONFIRMATION AND CONFIDENCE MEASURES

Monitoring calls to the VIOS system showed that many users have problems grasping the implicit confirmation strategy that is applied in this system. Things run smoothly as long as the recognizer makes no errors, but if recognition errors do occur, many callers are confused and fail to correct the system if the confirmation request is embedded in a prompt for additional information. Also, subjects do not know what to do if multiple items are confirmed in a single utterance, and only one or two are wrong.

To avoid these problems in the 98-ARISE system, we decided to explicitly confirm each information item in a separate question in the next turn. Experiments have shown that the explicit confirmation makes the dialogue more transparent and that errors are easier to correct [4]. The comparison of VIOS and the 98-ARISE system [6] showed that the explicit confirmation strategy does not increase the dialogue duration in seconds, because this strategy allows for very short prompts. However, the average number of turns does increase [4,6]. As a consequence, users indicated that the interaction becomes more tedious.

In the 99-ARISE system we apply a combination of implicit and explicit confirmation based on the confidence that the item is recognized correctly. If the confidence is high (i.e. if the risk that the caller should want to correct the system is low), implicit confirmation is applied; else, explicit confirmation is used [1]. Evaluations of this confirmation strategy [5] showed that implicit confirmation was applied in 56% of all the dialogue nodes where it could be used[1]. Of all implicit confirmations 6% actually

[1] The time of arrival/departure is always confirmed explicitly, because this is the last question to be answered before the travel advice is presented. At this point in the dialogue the cost of an incorrect implicit confirmation is too high to allow for implicit confirmation: an incorrect implicit confirmation results in a wrong travel advice.

contained incorrectly recognized information. In the successful dialogues, where the proportion of implicit confirmation is higher than the average, the number of turns decreased substantially thanks to the implicit confirmation. For the most simple scenario the modal number of turns in the slot-filling part of the dialogue was two turns lower than in the 98-version, where explicit confirmation was always applied: the modal number of turns decreased from six to four. In the two more complex scenarios, where one or two default values had to be changed, the modal number of turns decreased with three and four turns, respectively. In the 98-version changing a default took at least one turn. In the 99-system, if the correction is understood with a high confidence, changing defaults need not cost an extra turn.

In many of the successful dialogues the use of confidence measures essentially restores implicit confirmation. This way, in the successful dialogues with the 99-ARISE system the new confirmation strategy reduces the minimum and modal number of turns to the level of the VIOS system, where implicit verification is always applied. Moreover, the dialogues are shorter when measured in seconds.

3 MIXED INITIATIVE *VS.* SYSTEM DRIVEN INTERACTION

Both the VIOS and the ARISE system use a mixed initiative dialogue approach: the user can always provide the system with more information than (s)he is prompted for. A mixed initiative dialogue has several advantages compared to a system-guided dialogue: if a user can take the initiative, the dialogue becomes more natural. Furthermore, while novice users are guided through the dialogue and can provide the required information simply by answering all the questions, more experienced users who know what information the system needs to perform a database query, can provide that information immediately; this leads to shorter dialogues.

It turned out that in both the ARISE and the VIOS system the mixed initiative capabilities of the system were used only occasionally. Due to the directive questions the system asks (e.g., 'From where to where do you want to travel?' and 'Today?'), users do not spontaneously induce that they can take the initiative to provide information about the date or the arrival/departure time. Yet, in certain situations things are different. If a caller needs to negate a default assumption made by the system, the majority gives the correct value in the same turn as the negation. For instance, by default, the ARISE system assumes that the caller needs a connection for the same day[2]. This is expressed by the one word prompt 'Today?'. In two of the three scenarios that we used to evaluate the ARISE systems [4,5] users had to ask for a connection for the next day. Analyses of the dialogues show that subjects say just 'no' in only 15% of the cases. Eighty-five percent of the questions were answered by saying something like 'No, tomorrow'. In the 98-system this shortens the dialogue by one turn. In the 99-system, the dialogue can be shortened by two turns, if the item is recognized with a high confidence.

[2] Analysis of calls to the operator based service showed that most people ask for a connection on the same day.

4 EXCEPTIONS HANDLING

If the dialogue does not proceed, the dialogue manager has little means to diagnose the cause: is it because of recognition errors or because the caller does not understand what action the system expects? In the VIOS and ARISE systems continuous speech recognition (CSR) and natural language processing (NLP) can fail to find any useful information in an utterance. If that happens, the VIOS system just asks the caller to repeat the request. If no useful information is obtained after the third attempt, the system sends the call to an operator. The ARISE system applies a zooming strategy: on failure to extract any information from the user's utterance, it provides the user with hints on the options at that specific point in the dialogue, e.g. 'say the day that you want to travel' or 'say yes or no'. In the 98-system the question 'what did you say?' was asked before giving the hints, but this only helped if the user had not answered the original question at all. More often, this question resulted in an exact repetition of the user's utterance, which rarely solved the recognition problem. The 99-system proceeds immediately to give hints [5]. In 67% of all cases the user reformulated the answer to make it compatible with the hints. Most of the time the new answer was correctly understood. In the other 33% of the cases users did not reformulate their answer, often because their previous utterance already contained information compatible with the hints. A verbatim repetition of the previous answer helped only in a few cases.

In a way zooming is equivalent to starting with a mixed initiative dialogue strategy, and reverting to a system driven approach in case of problems. Thus, zooming does not constrain the caller unnecessarily, while it can profit from superior performance of the CSR during system driven phases of the dialogues, by restricting the lexicon and language model to the utterances that are reasonable given the question. An off-line experiment with training data from the VIOS system showed that making the lexicon, grammar, and language models dialogue-node dependent yields a relative improvement of 20% WER for the answers to the opening question. However, the results also showed that this improvement only holds for utterances that strictly reply to the question. When the user provides more information than (s)he is prompted for, the system performance decreases. Therefore, if zooming is used in conjunction with adaptive language (and acoustic) models, care must be taken to insure that the prompts avoid responses which are not in the restricted vocabulary and grammar; otherwise the strategy turns counterproductive [2].

5 NAVIGATION

The callers of the VIOS system can be divided in two categories: people who need factual information and those who rather seek assistance in planning a trip. Serving the second group requires that the system is able to negotiate. For the time being, that is beyond the state-of-the-art. To be at least somewhat helpful, the system must offer flexible navigation through the schedule information.

The VIOS system retrieves all connections in a time window determined by the user's query. It starts out by reading the earliest connection in the window; then, the caller is offered the option to get the next connection if (s)he is not satisfied with the first advice. This is repeated until no more connections are available in the window. Experience with similar systems in France and Germany suggests that this navigation strategy is adequate as long as no more than two or three connections are available. In the Dutch railway network some cities have very frequent connections, not all of which are equally convenient. This calls for a more flexible navigation.

Both ARISE systems start out by giving the single 'best' travel advice. In the navigation part of the 99-dialogue the user can ask for the previous or later train, for a connection with fewer changes, for information on platforms and directions of the trains, for the return trip, and for another connection. One can always ask for a repetition of the travel advice. In fact, we attempted to offer a speech driven version of the navigation through the information on the operator screen. The implementation of such an interface turned out to be difficult and a number of problems remain unsolved.

After the presentation of the 'best' travel advice, the system asks 'have you received sufficient information?'. If the user needs more information, (s)he can say directly which information (s)he wants. However, in our tests almost 50% of the users who needed extra information responded by simply saying 'no'. In that case an extreme form of zooming is applied: the user enters a menu. When the menu is entered, users almost always get the information they are looking for. However, if the user does not enter the menu, things go wrong more easily: in 23 of the 25 unsuccessful dialogues that concern the navigation part of the dialogue (the second and third tasks of scenarios II and III), the user did not enter the menu, because (s)he took the initiative and tried to indicate what information (s)he was looking for.

Most of the problems that occur in the navigation are caused by errors in the CSR and NLP. To a large extent these errors are artifacts of the scenarios, that seduce subjects to explain why they are seeking additional information. This leads to long utterances, with many out-of-vocabulary (OOV) words. But even if callers used in-vocabulary words, the recognition performance suffered from a lack of training data[3]. Nevertheless, from Table 1 it can be seen that the overall task completion rate in the navigation part of the dialogue is at the same level as observed in the slot-filling part of the dialogues.

Table 1 Dialogue success rate per task (99-system)
%Success = Success / (Total - Wrong data)

	Success	**Wrong data**	**Not compl.**	**Total**	**% Success**
Scenario I					
A → B	93	4	4	101	**96%**
Scenario II					
A → B	101	2	9	112	**92%**
Later	66	0	5	71	**93%**
B → A	88	1	6	95	**94%**
Scenario III					
A → B	81	2	6	89	**93%**
Less changes	62	0	6	68	**91%**
Platform info	58	1	8	67	**88%**
Total	**549**	**10**	**44**	**603**	**93%**

6 SHORTCOMINGS OF SCENARIOS

Both the ARISE and the VIOS system were evaluated by means of tests in which users had to complete predefined scenarios. The advantage of scenarios is that one can control the situation and test functions of the system that are not frequently used under normal circumstances. Furthermore, scenarios simplify performance evaluation, because it is known exactly whether the user got the 'correct' information. An additional advantage is that dialogues of different users and with different systems can be compared.

However, scenarios have considerable disadvantages as well. First, it is very difficult to present scenarios in such a way that the intention is unambiguous, while at the same time avoiding to suggest specific formulations. We tried to accomplish this by presenting the scenarios in a mix of text and graphics. Our tests [4,5] showed that graphics can only be used for very simple tasks. For more complicated tasks (e.g. ask for a connection with less changes) textual descriptions of the specific requirements had to be added. Text is easily repeated by the subjects, which induces problems for the speech recognizer, because the text often contains a lot of OOV words.

Second, Table 1 shows that the total number of dialogues for the second and third tasks of scenarios II and III is low compared to the total number of dialogues for the first task of the same scenarios. This indicates that a

[3] Language models could only be trained properly for the slot-filling part of the dialogue using 4,000+ recordings of interactions with the VIOS system. For the navigation part the relative frequencies of words and expressions had to be extrapolated from data of very few dialogues.

number of subjects did not even attempt to carry out the tasks that concern the navigation part of the dialogue. There are several ways in which this omission can be explained. During a lab test [4], where we observed the subjects, we found that some did not understand that they had to carry out these tasks, due to the way the scenario was described. Another explanation is that subjects are more willing to accept 'incorrect' information, because they do not really need the information they have to acquire according to the scenario. It may even happen that subjects do not notice that they get incorrect information. For instance, in scenario II users were encouraged to ask for a later connection, because the arrival time of the provided connection was 59 minutes earlier than the designated arrival time. However, some users did not even notice this difference, because probably they did not listen very carefully to the advice, since they did not really need the information. There is also a 'positive' explanation for the fact that some subjects accept information that is not strictly compatible with the instructions in the scenario. For instance, some subjects may have known that the least frequent train connections in the Netherlands provide one train each hour, so that they could easily infer that they could take another train, one hour later, without asking the system to provide this redundant information.

An analysis of the VIOS system has shown that the dialogue success rate for calls during the night hours (when the operator based version of the service is not available) is significantly higher than during office hours, because callers do not have an alternative way to obtain the information.

7 DISCUSSION & CONCLUSION

The comparison of the ARISE research system and the operational VIOS system has brought to light several problems related to the inability of the dialogue 'partners' to infer each other's mental model and intentions.

A large part of the problems with implicit confirmation can be solved by using confidence measures. Only confirming information that is very likely to be correct, while at the same time asking for additional information avoids confusion. Zooming in exceptions handling helps to avoid premature termination of the dialogue. In the slot-filling part of the dialogue, where only few hints are relevant at the given dialogue node, zooming is highly appreciated by the experts. This is not so in the navigation part of the dialogue: the number of options in the menu is so large that enumeration becomes tedious.

The caller has little or no means to infer the exact functionality of the system, especially in the navigation part of the dialogue[4]. Here we had to make decisions on the functionality, since the implementation of the full intelligence of an operator is far beyond the present capabilities. The only way to convey the limitations of the functionality is to explicitly enumerate the options, and hope that the caller will understand that e.g. the option to ask for the previous or following train means exactly this: one is only allowed the equivalent of a single cursor up/down action to scroll through a list of connections on a virtual screen; each cursor 'keystroke' is echoed by reading the corresponding connection. Also, we decided that one cannot navigate by mentioning a new departure/arrival time either, which was requested by quite a number of experts who were interviewed and naive subjects who were observed in the laboratory. But at the same time direct access to another connection is provided if that connection has fewer changes. The most promising solution to this problem that we can see today is to use a screen phone, which displays the list of possible connections. We are preparing experiments to investigate whether a visual display of summary data about the connections can facilitate navigation.

Also, the system has little means to infer the domain knowledge of the caller, nor the exact intention that (s)he tries to fulfill. This is only partly due to the fact that it is difficult to distinguish the utterances that are scrambled by the CSR/NLP from the utterances that do not contain relevant information. In some cases the caller may not really be interested in one exact connection, but rather in pattern information for some part of the day. Thus, if a caller fails to correct the system if it substitutes 'three p.m.' for 'four p.m.', it does not necessarily mean (s)he does not get adequate information. This makes it impossible for the experimenter to decide whether a dialogue is successful, if the 'hidden' intention of the user is the most important measure.

[4] Problems in the slot-filling part of the dialogue, for example with domain coverage, were avoided by using scenarios. If subjects are left free to invent queries, a small proportion will be about cities that do not have train stations.

REFERENCES

[1] Bouwman, G., J. Sturm, L. Boves (1999), "Incorporating Confidence Measures in the Dutch Train Timetable Information System Developed in the Arise Project", *Proc. ICASSP'99*, pp. I-493-496.

[2] Castagnieri, G., P. Baggia, M. Danieli (1998) "Field Trials of the Italian ARISE Train Timetable System", *Proc. IVTTA'98*, pp. 97-102.

[3] den Os, E., L. Boves, L. Lamel, P. Baggia (1999) "Overview of the ARISE project", *Proc. Eurospeech'99.*

[4] Sanderman, A., J. Sturm, E. den Os, L. Boves, A. Cremers (1998), "Evaluation of the Dutch Train Time Table Information System developed in the ARISE Project", *Proc. IVTTA'98,* pp. 91-96.

[5] Sturm, J., E. den Os, L. Boves (1999) "Dialogue Management in the Dutch Arise Train Timetable Information System", *Proc. Eurospeech'99.*

[6] Weegels, M. (1999), "Usability Evaluation of Voice-Operated Information Services: a Comparative Study of VIOS and ARISE", Dutch Priority Programme Language and Speech Technology, IPO Center for Research on User-System Interaction, TU Eindhoven, The Netherlands.